

Multi-Level Aggregation and Recursive Alignment Architecture for Efficient Parallel Inference Segmentation Network

Yanhua Zhang^a, Ke Zhang^{a,*}, Jingyu Wang^{a,b}, Yulin Wu^c, Wuwei Wang^d

^aSchool of Astronautics, Northwestern Polytechnical University, 127 West Youyi Road, Xi'an, 710072, China

^bSchool of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, 127 West Youyi Road, Xi'an, 710072, China

^cBeijing Aerospace Automatic Control Research Institute, Beijing, 100000, China

^dXi'an University of Posts and Telecommunications, Xi'an, 710072, China

Abstract

Real-time semantic segmentation is a crucial research for real-world applications. However, many methods lay particular emphasis on reducing the computational complexity and model size, while largely sacrificing the accuracy. To tackle this problem, we propose a parallel inference network customized for semantic segmentation tasks to achieve a good trade-off between speed and accuracy. We employ a shallow backbone to ensure real-time speed, and propose three core components to compensate for the reduced model capacity to improve accuracy. Specifically, we first design a dual-pyramidal path architecture (Multi-level Feature Aggregation Module, MFAM) to aggregate multi-level features from the encoder to each scale, providing hierarchical clues for subsequent spatial alignment and corresponding in-network inference. Then, we build Recursive Alignment Module (RAM) by combining the flow-based alignment module with recursive upsampling architecture for accurate spatial alignment between multi-scale feature maps with half the computational complexity of the straightforward alignment method. Finally, we perform independent parallel inference on the aligned features to obtain multi-scale scores, and adaptively fuse them through an attention-based Adaptive Scores Fusion Module (ASFAM) so that the final prediction can favor objects of multiple scales. Our framework shows a better balance between speed and accuracy than state-of-the-art real-time methods on Cityscapes and CamVid datasets. We also conducted systematic ablation studies to gain insight into our motivation and architectural design. Code is available at: <https://github.com/Yanhua-Zhang/MFARANet>.

Keywords:

Real-time semantic segmentation, parallel inference networks, multi-level feature aggregation, spatial alignment.

1. Introduction

Semantic segmentation, aiming to label every pixel in the image, is one of the basic tasks of computer vision. Recently, architectural advances in Deep Convolutional Neural Networks (DCNNs) for image classification have greatly facilitated other visual recognition tasks[1, 2, 3], including semantic segmentation[4, 5, 6]. Although DCNNs-based semantic segmentation methods have significantly improved the accuracy, they are difficult to be directly applied to real-world applications due to large model size or high complexity. Some applications, such as autonomous navigation or driver assistance systems, require segmentation models to maintain real-time inference speeds on large-size images to provide a wider field of view, placing a greater computational burden on the system. This makes real-time semantic segmentation a challenging research topic.

To achieve real-time inference speed on high-end Graphics Processing Unit (GPU) cards, several models reduce computational complexity by adopting convolution factorization or designing shallow architectures. Among them, ERFNet[7] decomposes the standard 3×3 convolution into 3×1 and 1×3 operations by asymmetric convolutions, which theoretically reduces the computation by a factor of 1.5. ICNet [8] uses a shallow network to obtain spatial details, and inputs downsampled images into a deeper network for semantic information. Although above methods significantly improve inference speed, their accuracy is largely reduced. For some real-world applications, the segmentation accuracy is also important.

Obtaining multi-scale features is one of the key points to improve segmentation accuracy, as objects appear at various scales in the field of view [9, 10, 11]. As shown in Fig. 2(a), image pyramid is a straightforward way to obtain hierarchical features. It scales the original image to multiple resolutions to generate a set of images, which are then independently fed into the same neural network to obtain multi-scale scores [9, 12, 13]. Due to its effectiveness in promoting performance, image pyramid-based multi-scale inference has become a common operation in the evaluation phase of accuracy-oriented methods [14, 15]. However, the high computational complexity and large memory

*Corresponding author

Email addresses: yanhuazhang@mail.nwpu.edu.cn (Yanhua Zhang), zhangke@nwpu.edu.cn (Ke Zhang), jywang@nwpu.edu.cn (Jingyu Wang), wuyulin@mail.nwpu.edu.cn (Yulin Wu), wangwuwei124@mail.nwpu.edu.cn (Wuwei Wang)

cost make it difficult to be directly used in real-world applications and end-to-end network training.

Compared with image pyramid, utilizing the inherent multi-scale property of DCNNs is a more efficient way to obtain hierarchical features [16], which is widely used by accuracy-oriented [17, 18, 19] and speed-oriented segmentation networks [20, 21, 22]. By stacking strided convolution layers, different stages of the DCNNs extract multi-level features with different receptive fields and spatial resolutions. In segmentation tasks, the most popular way to use this property is to aggregate multi-stage features to the highest-resolution level, and then make single-scale prediction upon the fused feature map [23, 22, 24]. For example, DFA [21] designs a dual-path decoder to aggregate multi-level features of a lightweight backbone, while Panoptic FPN [25] directly fuses the upsampled features through element-wise sum operation. However, such feature-level fusion will lead to less efficiency of hierarchical representation due to semantic gaps and spatial misalignment.

Instead of combining and predicting from the feature pipeline, another way is to do parallel inference at each layer of the encoder to generate multiple score maps, and then do a late fusion over them. This in-network parallel inference architecture proposed by FPN [16] (Fig. 2(b)) is widely used in object detection tasks [26, 27, 28, 29]. However, the adaptation of this architecture to semantic segmentation tasks has not been deeply explored. Some segmentation networks either simply treat multi-level predictions as an additional precision boosting strategy [30, 11, 31], or focus on using the coarse to fine strategy [32, 33, 34]. They perform predictions directly and independently at different stages of the encoder, thus generating coarse or fine score maps based on the semantic or local spatial detail information contained in the corresponding features. Here, we point out that the FPN architecture designed for object detection is not suitable to be directly applied to segmentation tasks for two main reasons: firstly, it is a suboptimal choice to obtain score maps at each level independently, because both high-level semantic information and low-level spatial details are important cues for segmentation [21, 22, 18]. The image pyramid (Fig. 2(a)) performs inference of each scale on multi-level features, while FPN lacks lower-level features in the top-most and mid-scale inference; Secondly, FPN for object detection does not need to consider the fusion of multi-scale pixel-level predictions, which is an important factor for fine-grained segmentation tasks. When fusing score maps, the above mentioned segmentation methods simply upsample low-resolution maps by bilinear interpolation, which is difficult to be used to recover the spatial misalignment caused by long-distance connections and repeated downsampling [35, 36, 37].

Differently, we design a parallel inference network customized for semantic segmentation tasks that can provide rich hierarchical features for the prediction at each level to obtain multi-scale scores as if obtained from image pyramid while being more efficient than it. To this end, we propose a Multi-level Feature Aggregation Module (MFAM) to aggregate hierarchical features of the encoder to each scale for further alignment and inference. Upon the features obtained from MFAM, we design a Recursive Alignment Module (RAM) to achieve spa-

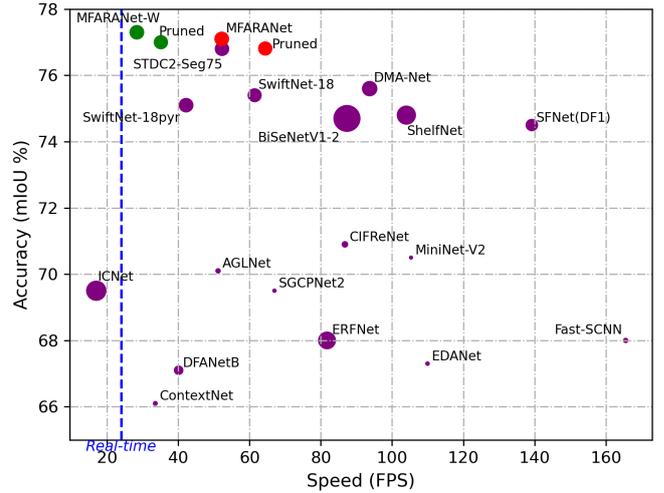


Figure 1: Inference speed and Params vs. mIoU accuracy on the Cityscapes test set. The size of the points indicates the model size. Purple points are representative real-time methods in Table. 9. The red and green points represent our MFARANet using 1024×1024 or the whole image (1024×2048) input for inference, respectively. 'Pruned' indicates the special network pruning method used for our parallel inference network, which is shown in Fig. 9 and discussed in Section. 4.2.8. All FPS are measured on a single GTX 3090 GPU with the image resolution on which the inference is performed to calculate the accuracy.

tial alignment between multi-scale score maps, which has half the computational complexity of the straightforward alignment method. By combining the flow-based alignment module with our proposed recursive upsampling architecture, the RAM employs intermediate features for stepwise alignment. Furthermore, by adopting the pixel-wise attention to fuse multi-scale scores in an adaptive manner, the proposed Adaptive Scores Fusion Module (ASFM) generates the final prediction that can benefit the segmentation of objects of multiple scales. It is worth mentioning that from the ablation experiments, we found a pruning method specific to this parallel inference architecture, that is, training on the entire network and removing certain scales during testing phrase. However, using network pruning to reduce model complexity is not the focus of this work, so it is only illustrated in Fig. 9 and discussed in Section. 4.2.8. Finally, to enhance network training, we propose the Multi-scale Joint Supervision (MJS) to add extra supervision at each scale of the predictions. It employs Cross-Entropy (CE) to calculate the loss of segmentation prediction and boundary prediction, and uses a regularization term to maintain consistency between them.

Our main contributions are summarized as follows.

1. We propose MFAM to fuse multi-level features in an information flowing manner through the dualpath architecture and the lateral connections. With this simple and efficient structure, we can aggregate low-level spatial details and high-level semantic information at each scale to alleviate the semantic gaps, facilitating subsequent spatial alignment and prediction.
2. To avoid spatial misalignment between scaled features, the RAM is designed to adopt intermediate features for step-

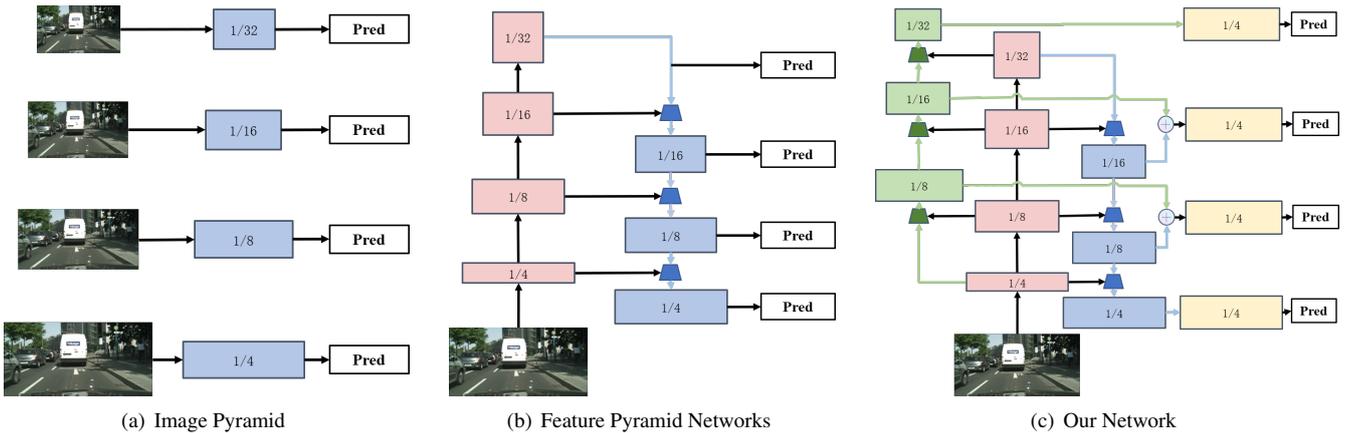


Figure 2: The comparison between different parallel inference networks. The boxes represent feature maps, and their length and number roughly reflect the relative spatial resolution between them. (a) Independently passing the scaled images through the whole segmentation network (omitted for simplicity) to obtain multi-scale predictions. (b) The in-network parallel inference architecture proposed by Feature Pyramid Networks (FPN) [16] for object detection tasks. (c) Our custom-designed parallel inference network for fine-grained real-time semantic segmentation tasks, which is much faster than image pyramid.

wise alignment, which is more accurate and faster than straightforward alignment. Furthermore, combined with MFAM, our RAM can align multi-scale score maps without being disturbed by semantic gaps.

3. Instead of fusing and inferring from the feature pipeline, we design a parallel inference network customized for semantic segmentation tasks to obtain multi-scale scores, and also build the pixel-wise attention based ASFM to adaptively fuse them to improve accuracy.
4. Finally, we use the MJS to jointly supervise segmentation prediction and boundary prediction at each scale to enhance feature representation and boost network training.

2. Related Works

There is a vast DCNNs-based literatures on accuracy-oriented or speed-oriented semantic segmentation. In the following, we mainly review the techniques of multi-scale/parallel inference, multi-level feature aggregation and feature alignment in semantic segmentation, which are related to our method.

2.1. Multi-scale/Parallel Inference

Due to the high intra-class and inter-class scale variations of objects, obtaining multi-scale features or predictions is one of the key points to improve accuracy in semantic segmentation [11, 10, 16]. Performing multi-scale inference on the image pyramid (Fig. 2(a)) is the most straightforward way to obtain multi-scale score maps. DeepLabv2 [38] extracts multiple score maps from a group rescaled original images through a shared network, and then performs cross scale fusion on them by maximum operations. Chen et al. [9] proposed using an attention mechanism to soft-weight each pixel between the score maps. Tao et al. [13] combined pixel-wise attention with the chain structure to fuse multi-scale predictions. At present, the multi-scale inference operation based on image pyramid has been widely used in the evaluation phase to boost segmentation accuracy [14, 15]. However, this method independently

passes multiple scaled images through the network, which exponentially increases the computational complexity and memory consumption, making it difficult to be applied to end-to-end network training and real-time required applications.

To promote efficiency, some networks perform parallel inference on multi-level feature maps of the encoder to generate multiple score maps. FCN8s [30] and Hypercolumns [11] are early attempts to fuse the top-most prediction with middle-layer high-resolution predictions to generate the final prediction, with the aim of enabling their networks to learn finer details. Similarly, DAG-RNN [31] and DeepLab [39] directly employ this multi-level parallel inference method as an additional performance boosting strategy. NDNNet [40] uses convolutional layers to perform weighted fusion of multi-scale scores, and adapts separable convolutions to build a lightweight FCN8 to achieve real-time inference speed. GCN [32], CGBNet [33], and SABNet [34] all follow a similar network structure design pipeline and adopt a coarse to fine strategy, which utilizes high-resolution fine predictions from middle layers to gradually refine the topmost low-resolution coarse prediction.

It is a suboptimal choice to infer independently on single-level features of the encoder to obtain coarse or fine score maps, because multi-level features together provide rich cues for segmentation. Differently, we propose MFAM to aggregate multi-level features of the encoder to each scale, providing hierarchical information for further prediction. Besides, by using the RAM, we solve the spatial misalignment problem in the fusion process of multi-scale scores, which is ignored by the above methods.

2.2. Multi-level Feature Aggregation

As an effective method to improve accuracy, multi-level feature aggregation is used by both accuracy-oriented [25, 19] and speed-oriented segmentation methods [21, 41, 22]. The top-down path is used by FPN [16] (Fig. 2(b)) to aggregate other levels of features to the lowest-level feature in an information

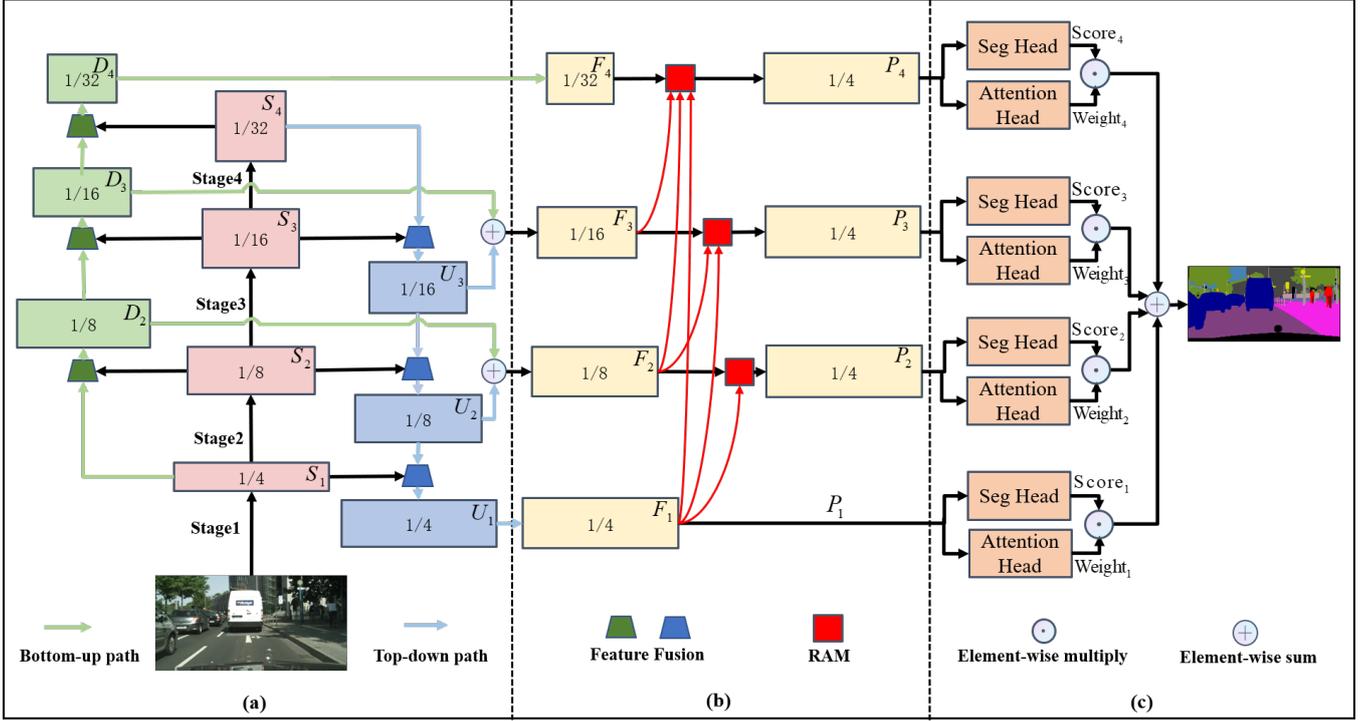


Figure 3: Overall architecture of our approach. (a) The details of MFAM. S_i , D_i and U_i represent multi-level features from different stages of ResNet-18, Bottom-up path and Top-down path, respectively. (b) The process of aligning low-resolution features with the highest-resolution feature F_1 . The detailed structure of RAM is illustrated in Fig. 4(b). F_i indicates the i -th scale feature obtained from MFAM, and P_i denotes the corresponding aligned feature. (c) The architecture of ASFM. “Seg Head” and “Attention Head” represent the modules for obtaining the score map ($Score_i$) and weight map ($Weight_i$) of the i -th scale, which are shown in detail in Fig. 6.

flow manner, in which mid-level features are fused through lateral connections. Based on FPN-like architecture, lots of methods directly adopt element-wise sum [25] or channel concatenation [42, 19] to fuse multi-level features. Among the sophisticated methods, feature reuse is a commonly utilized mechanism. PANet [23] and SGCPNet [43] add a bottom-up path to integrate the middle layer features of the top-down path. In contrast, DFANet [21] and RGPNet [24] first employ the bottom-up path to aggregate features from the encoder, and then adopt a top-down path to reuse features of the bottom-up path. ShelfNet [41] uses three overlapping paths to reuse features from the previous path. For approaches different from feature reuse, Lin et al. [17] proposed an intertwining architecture to exchange information between adjacent scales in a bidirectional fashion. ZigZagNet [44] employs dense connections to blend feature maps between top-down and bottom-up networks. Weng et al. [22] used a long-distance skip connection to fuse the lowest-level feature with the middle-level features in the top-down path for real-time segmentation. All the above methods integrate multi-level features into the highest-resolution feature to obtain a single score map, while our MFAM aggregates multi-level features to each scale to generate multi-scale score maps.

2.3. Feature Alignment/Upsampling

DCNNs need to enlarge the receptive field through repeated max-pooling or stride convolution, so the high-level features

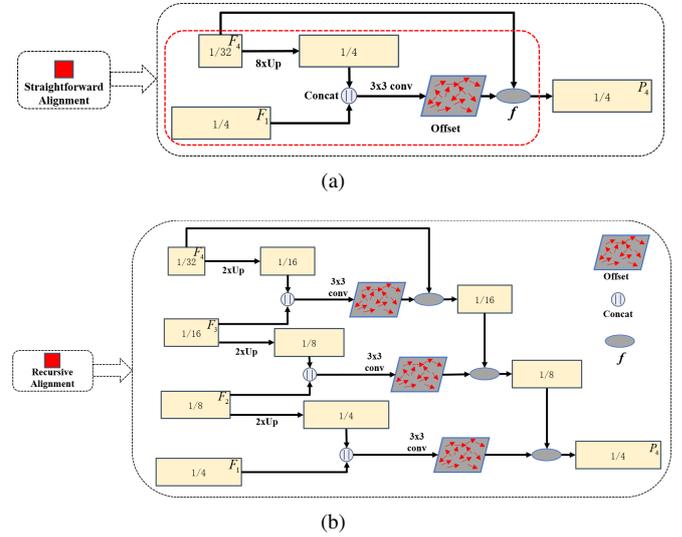


Figure 4: The comparison between Straightforward Alignment and Recursive Alignment. The red dashed box indicates the flow based alignment module. Up represents bilinear interpolation. Concat denotes the channel concatenation operation. f is the alignment function in Eq. 4. (a) Straightforward Alignment Module. (b) Our proposed Recursive Alignment Module (RAM).

of the encoder and the final predictions are inevitably down-sampled. Bilinear interpolation is widely used to upsample feature/score maps [38, 10, 12], as it is simply and fast. However, it is oversimple and data-independent, making it difficult to align

multi-level features. To avoid the spatial misalignment, Noh et al. [45] adopted unpooling layers and deconvolution layers to build the decoder to recover the details of low-resolution feature maps. To ensure efficiency, SegNet [46] uses the pooling positions stored in the encoder to achieve nonlinear upsampling of the corresponding level features of the decoder. Recently, GUN [35] proposes a transformation module to learn 2D offsets to guide upsampling between two different resolution features for real-time segmentation. AlignSeg [47], FaPN [48] and SFNet [42] share a similar motivation and all propose using flow-based alignment module to estimate 2D transformation offsets to implement learnable interpolation. The above flow-based alignment methods fuse features between different levels of the encoder, so they need to bridge the semantic gap when performing spatial alignment. The semantic gap between multi-level features may mislead the transform offset learning in the alignment module. Furthermore, they are used for adjacent features without large resolution differences, while our RAM is specifically designed for aligning long-distance feature maps.

3. Methods

3.1. Overview

As shown in Fig. 3, three complementary modules are proposed to build our parallel inference network: MFAM, RAM and ASFM. For MFAM, we build it through top-down and bottom-up paths, as well as the lateral connections for information exchange between paths. Such a structure has three advantages. Firstly, compared to long-range skip connections (Fig. 5), it can aggregate long-distance features in an information flow manner to narrow the semantic gap. Secondly, by aggregating hierarchical features of the encoder to each scale, this architecture can provide both high-level semantic information and low-level spatial details for the prediction of the corresponding scale. Thirdly, since each feature map from MFAM aggregates multi-level features of the same encoder, we have reason to believe that the semantic gap between the output features is largely narrowed. This will benefit the transform offset learning in the following spatial alignment module. Upon the multi-scale features obtained from MFAM, we build four independent parallel paths to perform multi-scale inference. Before ASFM, we first design RAM to combine flow-based alignment module with the recursive upsampling architecture to address the spatial misalignment between multi-scale score maps. In the end, we propose ASFM to do adaptive fusion over multi-scale scores, and the training is carried out by our customized supervision function MJS.

3.2. Multi-level Feature Aggregation Module (MFAM)

We start with the encoder to describe how to build our MFAM. There are two main reasons for choosing ResNet-18 as the backbone to obtain multi-level features. First, it is a good fit for fine-tuning since the pre-trained parameters are publicly available. Our network can benefit from knowledge transfer by loading their weights trained on imageNet. Second, due to the

moderate depth and residual structure, it promotes training efficiency and has low model complexity, which can be used to build real-time segmentation networks. The ResNet-18 has four stages (a.k.a., blocks), which extract multi-level feature maps with spatial resolutions of 1/4, 1/8, 1/16, and 1/32 of the input image ($S_1 \sim S_4$ in Fig. 3(a)). Additionally, we use two atrous convolutional layers [38] with dilation rate of 2 and 4 to replace the stride convolutions in the last two stages of the backbone to enlarge receptive field. Differently, we do not change the stride length of stride convolutions from 2 to 1 as done in [38, 10], thus increasing the receptive field of the last two stages without changing the feature size (still 1/16 and 1/32 of the input image). This can significantly increase accuracy with a slight decrease in inference speed (Table. 1, native FPN vs. improved FPN).

Upon multi-level features of the encoder, we flow the topmost-level feature to the lowest-level feature through the top-down path. In this process, we utilize feature fusion modules to aggregate mid-level features and pass them to the lower-level features, so that the outputs of top-down path fuse higher-level features (e.g., U_2 aggregates $S_2 \sim S_4$). It is worth noting that all the features from the encoder use a 1×1 convolutional layer to adjust the channel dimension before inputting the top-down path or bottom-up path, which is omitted in Fig. 3(a) for brevity. For the feature fusion modules in top-down path, we adopt a simple and general design to ensure efficiency. We use bilinear interpolation to upsample higher-level features, and then use element-wise addition followed by a 3×3 convolutional layer to fuse features. This process can be expressed mathematically as:

$$U_n = \text{conv}_{3 \times 3}(\text{Up}(U_{n+1}) + \text{conv}_{1 \times 1}(S_n)), n = 3, 2, 1. \quad (1)$$

Here, we directly get U_4 from S_4 : $U_4 = \text{conv}_{1 \times 1}(S_4)$. $\text{conv}_{1 \times 1}$ represents a 1×1 convolutional layer for feature channel adjustment. $\text{conv}_{3 \times 3}$ is a 3×3 convolutional layer for feature fusion. Up means bilinear upsampling. Similarly, we construct a bottom-up path to flow the lowest-level feature in the encoder to the topmost feature, so that the resulting features aggregate the lower-level features. We still employ a simple feature fusion module to fuse adjacent-level features, but we use bilinear interpolation to downsample features instead of upsampling. This can be mathematically expressed as:

$$D_{n+1} = \text{conv}_{3 \times 3}(\text{Down}(D_n) + \text{conv}_{1 \times 1}(S_{n+1})), n = 1, 2, 3, \quad (2)$$

where $D_1 = \text{conv}_{1 \times 1}(S_1)$; Down means bilinear downsample. By using the above two paths, the lowest-level feature (U_1) in the top-down path and the topmost-level feature (D_4) in the bottom-up path aggregate the features of all levels in the encoder, but it has not been done in the mid-scale features (e.g., D_2 fuses S_1, S_2 but lacks S_3, S_4). Therefore, we add two lateral connections between two paths to flow information between them, so that the mid-scale features (F_2 and F_3) could contain full-level features. For example, D_3 aggregates S_1, S_2 and S_3 , while U_3 includes S_4 . Then, we can obtain F_3 that fuses $S_1 \sim S_4$ through the lateral connection between D_3 and

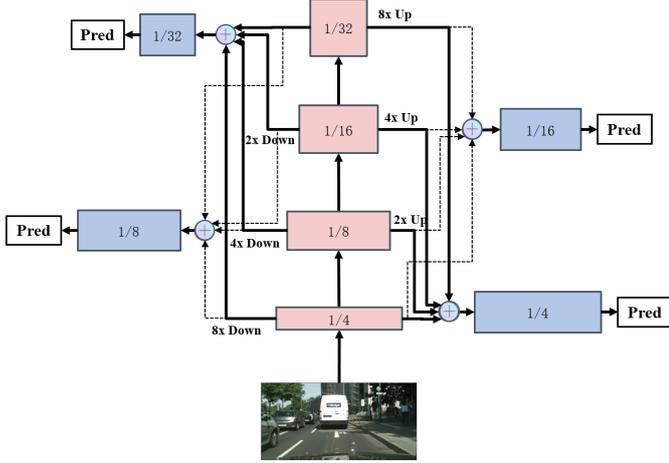


Figure 5: Long-range Skip Connection based Aggregation architecture (our initial design).

U_3 . This can be expressed as:

$$F_n = \text{conv}_{3 \times 3}(D_n + U_n), n = 2, 3. \quad (3)$$

We compare our MFAM with FPN-like architecture [16, 25, 19, 42] (Fig. 2(b)) and Long-range Skip Connection based Aggregation (LSCA) architecture illustrated in Fig. 5 to give readers a clearer understanding of the motivations of our design. FPN-like architecture utilizes the top-down path with lateral connections to obtain high-level semantic features at all stages, while its topmost and mid-scale features lack low-level information. The LSCA is the one we originally designed for fusing multi-level features. Although it could fuse multi-level features in an explicit manner, using the long skip connections between long-distance features (e.g., S_4 to S_1) for fusion will lead to a larger semantic gap and misalignment than fusion between adjacent features. Quantitative comparisons are given in Table. 2.

3.3. Recursive Alignment Module (RAM)

Inspired by [47, 48, 42], we adopt a flow-based alignment module in RAM to learn 2D transformation offsets to guide the upsampling between features with different resolutions. However, we have noticeable differences from above methods. Firstly, the aim of our RAM is to achieve spatial alignment between score maps by aligning features before the segmentation head, not for feature fusion. Secondly, the previous MFAM is complementary to our RAM, which can narrow the semantic gap between multi-scale features to prevent the offset learning from being misguided. Thirdly, we align features in a stepwise manner specifically designed for long-distance feature maps, while the above methods use their flow-based alignment module for adjacent features without large resolution differences.

As illustrated in the red dashed box in Fig. 4(a), the flow-based alignment module can be mathematically formulated as:

$$\begin{aligned} \Delta &= \text{conv}_{3 \times 3}(\text{concat}(F_h, \text{Up}(F_l))), \\ \hat{F}_l &= f(F_l, \Delta). \end{aligned} \quad (4)$$

Here, we use a 3×3 convolutional layer to learn the offset (Δ) between the high-resolution feature (F_h) and low-resolution feature (F_l). concat represents channel concatenation operation and Up refers to the bilinear interpolation upsampling. f denotes the alignment function for aligning feature with the learned offsets, and \hat{F}_l is the aligned feature.

To achieve our recursive alignment, we adopt intermediate features to learn offsets. Specifically, we learn the offsets between F_1 and F_2 , F_2 and F_3 , F_3 and F_4 , and then use these offsets to align the features in a step-wise manner. Therefore, the function for calculating the offset in Eq. 4 needs to be rewritten as:

$$\Delta_n = \text{conv}_{3 \times 3}(\text{concat}(F_n, \text{Up}(F_{n+1}))), n = 1, 2, 3. \quad (5)$$

After obtaining the offsets, we align low-resolution features (F_4, F_3, F_2) to F_1 respectively. It is worth noting that these offsets only need to be calculated once and can be shared by all three alignment processes. Fig. 4(b) shows an example of aligning F_4 to F_1 : we first use Δ_3 to align F_4 to F_3 , then use Δ_2 to align the output to F_2 , and finally use Δ_1 to align the previous output to F_1 . The aforementioned processes can be mathematically written as:

$$P_n = \left\{ \hat{F}_i = f(\hat{F}_{i+1}, \Delta_i), i = n-1, \dots, 1 \right\}_{n=2,3,4}, \quad (6)$$

where the initial value $\hat{F}_n = F_n$; P_n is the final aligned feature map corresponding to F_n ; $\{\cdot\}$ represents the recursive process. As for f , AlignSeg [47] modifies bilinear interpolation as their alignment function, while SFNet [42] and FaPN [48] employ the differentiable bi-linear sampling mechanism and deformable convolutions [49], respectively. With the demand for keeping efficiency, we directly use the alignment function proposed in SFNet. After having calculated Δ_n , the warped grid between F_{n+1} and F_n can be calculated as:

$$\text{warp}_n = \frac{g_n + \Delta_n(g_n)}{2}, \quad (7)$$

where g_n represents each position in spatial grid. In the end, we adopt the bi-linear sampling function to align features:

$$\hat{F}_{n+1} = f(F_{n+1}, \Delta) = \sum_{i \in \mathcal{N}(\text{warp}_n)} w_i F_{n+1}(i), \quad (8)$$

where $\mathcal{N}(\text{warp}_n)$ indicates the neighbors of warp_n in F_{n+1} , and w_i is the kernel weight estimated by warped grid.

In Fig. 4, we take the alignment between F_4 and F_1 as an example to compare our proposed RAM with the straightforward alignment method. As shown in Fig. 4, the straightforward method puts F_4 and F_1 into the flow-based alignment module to learn the corresponding offset, and then use the alignment function to directly align F_4 to F_1 without using intermediate features. Compared with the straightforward alignment, our designed RAM is not only more accurate but also more efficient. Here, we give computational complexity analysis. Since the complexity of other operations in RAM and the straightforward alignment method is negligible compared to the 3×3 convolutional operation, the overall complexity can be estimated by

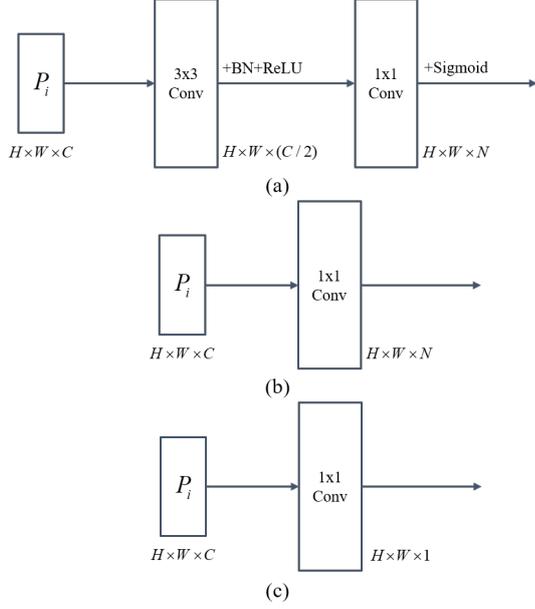


Figure 6: Detailed design of Attention Head, Segmentation Head and Boundary Head. P_i is the aligned feature at i -th scale. BN denotes the batch normalization. $H \times W \times C$ represents the height, width and channels of the input tensor. N means the number of classes. (a) Attention Head. (b) Segmentation Head. (c) Boundary Head.

3×3 convolutional operations. Therefore, the complexity of computing the offset between F_4 and F_3 can be expressed as $W \times H \times C_{in} \times C_{out} \times 3 \times 3$, which can be simplified as $9WHC^2$. Here, H and W denote the height and width of F_3 , and C_{in} and C_{out} represent the channels of input and output features of the convolutional layer. Since the height and width of F_1 are increased by 4 times compared to F_3 , the complexity of inputting upsampled F_4 and F_1 into the 3×3 convolutional layer can be expressed as $4W \times 4H \times C_{in} \times C_{out} \times 3 \times 3$. Ultimately, the ratio between the overall complexity of RAM and the straight-forward alignment method is: $(9WHC^2 + 4 \times 9WHC^2 + 16 \times 9WHC^2) / (3 \times 16 \times 9WHC^2) = 7/16$. In terms of accuracy, the quantitative comparison and corresponding analyses are given in Section. 4.2.3.

3.4. Adaptive Scores Fusion Module

Predictions at certain scales are better at solving objects of a specific scale. In detail, large-scale features are more suitable for segmenting small-scale objects due to their large resolution and small receptive field, while small-scale features with larger receptive field and rich context information are better at segmenting large-scale objects. Therefore, we design ASFM to fuse multi-scale scores to favor objects of various scales. Inspired by [13, 9], we employ the pixel-wise attention mechanism for adaptive fusion. Differently, we adopt this mechanism to fuse multiple scores obtained from our designed single pass network rather than image pyramids, greatly prompting efficiency. Besides, by combining with RAM, we avoid the spatial misalignment between multi-resolution score maps, which is not addressed by previous methods. As shown in Fig. 3(c), this mechanism can be easily implemented in our network. We

use a 1×1 convolutional layer to build the segmentation heads (Fig. 6(b)) to obtain score maps, and use a 3×3 convolutional layer followed by a 1×1 convolutional layer to build the attention head (Fig. 6(a)) to obtain the weight map at each scale. This process can be expressed mathematically as:

$$\begin{aligned} \text{Score}_n &= \text{conv}_{1 \times 1}(P_n), \\ \text{Weight}_n &= \text{conv}_{1 \times 1}(\text{conv}_{3 \times 3}(P_n)). \end{aligned} \quad (9)$$

Then, ASFM uses the weight maps to linearly fuse multi-scale scores to obtain the final segmentation result:

$$\text{Output} = \sum_{n=1}^4 \text{Scale}_n \odot \text{Weight}_n, \quad (10)$$

where \odot is element-wise multiplication.

3.5. Multi-scale Joint Supervision

To further improve segmentation accuracy, we propose the Multi-scale Joint Supervision (MJS) as our booster training strategy. Additional supervision during the training phase is a cheap way to improve segmentation accuracy as it can be discarded during the inference phase. Inspired by [50, 51], we supervise both segmentation and boundary prediction at each scale to enhance feature representation. Similar to the segmentation head, we directly add a boundary head (Fig. 6(c)) on the feature of each scale to produce the s -th scale boundary map f_b^s . The ground-truth semantic labels \hat{y} are passed through a Sobelfilter [52] to produce a binary boundary label \hat{y}_b . At the s -th scale, our MJS uses balance Binary Cross-Entropy (BCE) loss on the binary boundary prediction f_b^s and standard Cross-Entropy (CE) loss on segmentation prediction f^s . Besides, we adopt a regularization term proposed by GSCNN [50] in our loss function, which is used to keep consistency between boundary prediction f_b^s and segmentation prediction $f^s = p^s(y|P_s)$:

$$\mathcal{L}_{reg}^s = \sum_{c,i} \mathbb{1}_{f_b^s}^i [-\hat{y}_i^c \log p^s(y_i^c | P_s)], \quad (11)$$

where i and c run over all image pixels and C semantic classes. $\mathbb{1}_{f_b^s} = \{1 : f_b^s > \text{ths}\}$ is the indicator function, and the threshold (ths) is set to 0.8 according to GSCNN. The final MJS can be written as:

$$\mathcal{L} = \sum_{s=1}^4 (\lambda_1 \mathcal{L}_{BCE}(f_b^s, \hat{y}_b) + \lambda_2 \mathcal{L}_{CE}(f^s, \hat{y}) + \lambda_3 \mathcal{L}_{reg}^s), \quad (12)$$

where $\lambda_1, \lambda_2, \lambda_3$ are three hyperparameters that control the weights between losses, and we set them to 0.3, 1, 0.1, respectively, according to the settings in GSCNN. In addition, Online Hard Example Mining (OHEM) [53] is applied to mitigate overfitting and data imbalance. Following [54], we set the threshold for selecting hard pixels as 0.7, and keep at least 100,000 pixels within each minibatch.

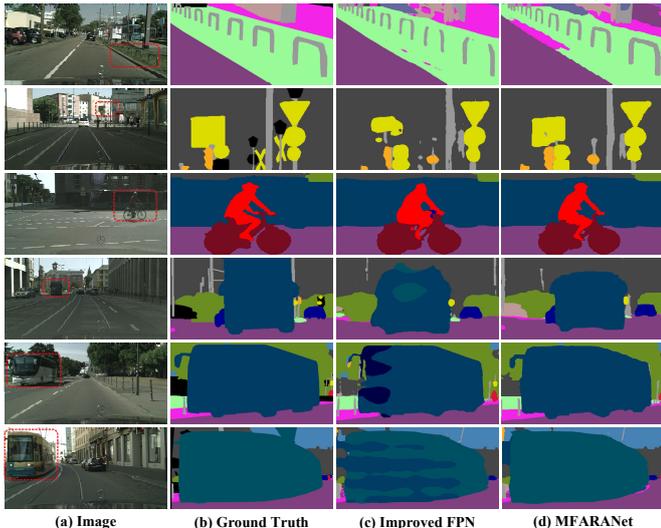


Figure 7: Qualitative results on the valuation set of Cityscapes. Compared with the strong baseline, our method shows significant improvement on small-scale (pole, traffic sign), middle-scale (rider, bus) and large-scale (bus, train) objects.

4. Experiments

In this section, we first introduce the datasets, metrics, and implementation details of our experiments. Then, we conduct ablation experiments on the Cityscapes validation set. Finally, comprehensive experiments are conducted on three benchmarks: Cityscapes, CamVid, and PASCAL-Context to compare with state-of-the-art methods.

4.1. Training Protocol

1) Dataset: **Cityscapes [55]**: It is a widely used dataset in semantic segmentation, focusing on semantic understanding of urban street scenes. It contains 20K coarsely annotated images and 5K finely annotated images, and we only use the finely annotated images to train and validate our network. The finely annotated images are split into 2, 975, 500 and 1, 525 subsets for training, validation and testing, respectively. 19 semantic classes are picked for the segmentation task. The dataset samples have a high resolution of 1, 024 × 2, 048, which is challenging for real-time semantic segmentation. We randomly crop all images to 1, 024 × 1, 024 for training.

CamVid [56]: It is another road scene dataset, which contains 701 images with 960 × 720 resolution. According to [57], 367, 101 and 233 images are selected for training, validation and testing, respectively. We use 11 semantic classes for segmentation task and employ random cropping of 640 × 640 crop size during training.

PASCAL-Context [58]: It has 10103 images, out of which 4998 images are used for training. Each image has approximately 375 × 500 pixels, which are randomly cropped to 480 × 480 for training. Following [33, 59, 60, 61], we only consider the most frequent 59 classes in the dataset for evaluation.

2) Metrics: We use the most commonly used class-wise mean Intersection over Union (mIoU) to evaluate segmentation accu-

Table 1: Comparison with the baseline method. MSL: Multi-Stage Loss. MJS: Multi-scale Joint Supervision. OHEM: Online Hard Example Mining.

Name	MSL	MJS + OHEM	mIoU (%)	Speed (FPS)	FLOPs (G)	Params (M)
native FPN	✓		72.3	54.2	112.1	12.5
improved FPN	✓		74.7	50.8	112.1	12.5
MFARANet		✓	77.0	52.2	101.6	13.0
MFARANet		✓	78.2	52.2	101.6	13.0

Table 2: Comparison with different multi-level feature aggregation methods. FPN-like: the FPN-like Network shown in Fig. 2(b). LSCA: the Long-range Skip Connection based Aggregation Architecture in Fig. 5.

FPN-like	LSCA	MFAM	RAM	ASF	MSL	mIoU (%)	Speed (FPS)	FLOPs (G)	Params (M)
✓			✓	✓	✓	76.3	57.5	95.8	12.3
	✓		✓	✓	✓	75.4	52.5	96.0	12.4
		✓	✓	✓	✓	77.0	52.2	101.6	13.0

racy. The formula is:

$$mIoU = \frac{1}{C} \sum_{i=1}^C \frac{N_{ii}}{\sum_{j=1}^C N_{ij} + \sum_{j=1}^C N_{ji} - N_{ii}}, \quad (13)$$

where C represents total categories; N_{ij} is the number of pixels of class i being predicted to be the class j ; $\sum_{j=1}^C N_{ij}$ is the number of pixels of class i ; $\sum_{j=1}^C N_{ji}$ refers to the false negative.

To evaluate the model complexity, we use the number of parameters (Params) to measure the size of networks and the Floating Point Operations (FLOPs) to calculate computational complexity. Besides, since the FLOPs poorly corresponds with the actual processing time on GPU platforms, we additionally use Frames Per Second (FPS) to measure execution speed. Following [62, 12], we exclude the batch normalization layers when measuring FPS, as these layers can be fused with preceding convolutional layers during inference.

3) Implementation Details: We conduct experiments based on PyTorch 1.8.1. The inference speed is measured on one NVIDIA GeForce GTX 3090 with the CUDA 11.1 and CUDNN 8.0. Following the prior protocol [10], we employ the “poly” learning rate policy by $1 - \left(\frac{iter}{max_iter}\right)^{power}$, in which we set base learning rate to 0.005 for Cityscapes and 0.001 for CamVid and PASCAL-Context. The value of $power$ is 0.9. During training, the network is optimized using the Stochastic Gradient Descent (SGD) algorithm, of which the momentum is 0.9 and weight decay is $5e-4$. For data augmentation, we employ random horizontal flip, random rotation from -10 to 10 degrees and random scale from 0.5 to 2.0. The batch size is 14 for Cityscapes and 16 for CamVid and PASCAL-Context. The epochs for Cityscapes, CamVid and PASCAL-Context are set to 200, 120, and 100, respectively. For inference, we do not adopt time-consuming evaluation tricks (e.g., horizontal flipping and multi-scale testing) to improve accuracy. Instead, we use the whole picture as input to compute mIoU unless explicitly mentioned.

4.2. Ablation Studies

In this section, we conduct the ablation experiments to verify the effectiveness of each module in our MFARANet. In the

Table 3: Ablation study on RAM. Bilinear: bilinear interpolation. SA: Straightforward Alignment.

MFAM	Bilinear	SA	RAM	ASFM	MSL	mIoU (%)	Speed (FPS)	FLOPs (G)	Params (M)
✓	✓			✓	✓	75.5	55.2	100.6	12.9
✓		✓		✓	✓	76.5	49.6	102.7	13.0
✓			✓	✓	✓	77.0	52.2	101.6	13.0

Table 4: Comparison between three multi-scale scores fusion methods: Average, Maximum, and ASFM.

MFAM	RAM	Max	Average	ASFM	MSL	mIoU (%)	Speed (FPS)	FLOPs (G)	Params (M)
✓	✓	✓			✓	75.8	57.5	83.3	12.7
✓	✓		✓		✓	75.9	57.0	83.3	12.7
✓	✓			✓	✓	77.0	52.2	101.6	13.0

following, we train our models on the training set of Cityscapes and evaluate on the validation set. FLOPs and FPS are computed on a single GPU with an input size of 1024×1024 .

4.2.1. Comparison with the Baseline Methods

Here, we choose a modified version of FPN network adapted for segmentation tasks as the baseline method (named as 'native FPN' in Table. 1), which aggregates the multi-level outputs of FPN to the highest resolution feature for single-scale prediction [25, 19, 42]. For a fair comparison, we also adopt two atrous convolutional layers to the backbone of the baseline, following the way described in Section. 3.2. Furthermore, we add three additional segmentation heads at different stages of the baseline to compute auxiliary loss to boost network training, which is named Multi-Stage Loss (MSL). Notably, by inferring on multi-scale features, our MFARANet can calculate MSL without adding additional segmentation head. As shown in Table. 1, through the above operations, we improve the accuracy of the native FPN from 72.3% to 74.7% of the improved FPN with a slight decrease in inference speed. Using the same backbone and MSL, our method can achieve 77.0% mIoU, which is 2.3% higher than the strong baseline. At the same time, the computational complexity is reduced by 10.5 GFLOPs, achieving a faster inference speed. Furthermore, by combining MJS and OHM to boost the training process, our method further improves the accuracy to 78.2%, which is 5.9% and 3.5% higher than native FPN and improved FPN, respectively. Qualitative segmentation comparison results between our method and the baseline on Cityscapes validation set are shown in Fig. 7. Our method shows improvement on multi-scale objects.

4.2.2. Ablation Study for MFAM

To demonstrate the superiority of MFAM, we compare it with the other two feature aggregation methods: FPN-like architecture (Fig. 2(b)) and LSCA (Fig. 5). In Table. 2, we replace the MFAM in our network with FPN and LSCA, respectively, and all other components remain unchanged. From the results, our method improves the accuracy by 0.7% and 1.6% compared to FPN and LSCA, respectively, with less than 6 GFLOPs increase in computational complexity. We attribute the superiority of our MFAM as two aspects: a) The topmost and mid-scale features

Table 5: Ablation study on importance of each component. "-": Indicates that the corresponding experiment directly utilizes the features of the encoder for alignment and prediction. Bilinear: bilinear interpolation. Concat: the channel concatenation operation for feature fusion.

MFAM	RAM	ASFM	MSL	mIoU (%)	Speed (FPS)	FLOPs (G)	Params (M)
-	✓	✓	✓	74.4	61.3	84.0	11.8
✓	Bilinear	✓	✓	75.5	55.2	100.6	12.9
✓	✓	concat	✓	76.4	44.5	118.8	13.3
✓	✓	✓	✓	77.0	52.2	101.6	13.0

Table 6: Ablation Study for booster training strategy. OHM: Online Hard Example Mining.

Name	MSL	MJS	OHM	mIoU(%)
	✓			77.0
MFARANet		✓		76.1
	✓		✓	77.1
		✓	✓	78.2

of FPN lack low-level features, while each feature from MFAM aggregates full-level features of the encoder. b) We aggregate long-distance features in an information flow manner, which narrows the semantic gap compared to using long-range skip connections.

4.2.3. Ablation Study for RAM

We propose RAM to align scaled features and carry out several ablation studies to reveal its effectiveness. First, we utilize bilinear interpolation to replace RAM for upsampling, and the first row in Table. 3 has a decrease in accuracy of 1.5% compared to the third row. This shows that spatial misalignment is one of the key issues for multi-scale scores fusion and can be efficiently solved by our RAM. Second, we compare Straightforward Alignment (Fig. 4(a)) with our proposed Recursive Alignment (Fig. 4(b)), and our RAM improves the accuracy and inference speed by 0.5% and 2.6 FPS, respectively. We attribute this to adopting the stepwise alignment to replace direct alignment between long-distance features. Adjacent features have less spatial misalignment due to smaller scale differences compared to long-range features, leading to more accurate offsets learning. Computational complexity analysis has been given in Section. 3.3.

4.2.4. Ablation Study for ASFM

We design ASFM to fuse multi-scale scores in an adaptive manner. Three fusion methods are compared in Table. 4: Average, Maximum (Max), and ASFM. When using Average or Max operations, the score maps at each scale are either equally important or sparsely selected, making them a sub-optimal choice. ASFM improves the segmentation accuracy by more than 1.0% over the other two methods, demonstrating the effectiveness of this adaptive fusion strategy. Compared with the promotion of accuracy, the increase in computational complexity is acceptable, as the inference speed is only reduced by around 5 FPS and remains above the real-time speed.

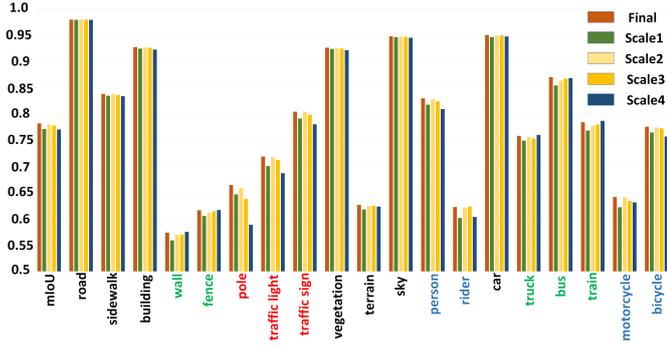


Figure 8: Comparison of segmentation accuracy on different classes at each scale and the final prediction. Scale1 ~ Scale4 indicate the corresponding predictions of Score₁ ~ Score₄ in Fig. 3(c). "Final" represents the fusion result of Scale1 ~ Scale4 using ASFM. Categories marked in red, blue, and green are small-scale, middle-scale, and large-scale objects, respectively.

Table 7: Ablation for Multi-scale predictions. "-": The corresponding score map and weight map are discarded during **training and inference** phase. The proposed MJS + OHEM is used to supervise the training of each experiment.

Scale1	Scale2	Scale3	Scale4	mIoU (%)	Speed (FPS)	FLOPs (G)	Params (M)
✓	-	-	-	76.93			
-	✓	-	-	77.20	56.9	83.3	12.7
-	-	✓	-	76.72			
-	-	-	✓	76.51			
-	✓	✓	-	77.67			
✓	-	✓	✓	77.12	53.2	97.0	12.9
✓	✓	-	✓	77.61			
✓	✓	✓	-	77.57			
✓	✓	✓	✓	78.16	52.2	101.6	13.0

4.2.5. Importance of each Module

In Table. 5, we conduct three experiments to reveal the importance of feature aggregation, spatial alignment and adaptive score-level fusion. In the first row, we directly input the features from each level of the encoder into subsequent modules. This results in a significant reduction in accuracy of 2.6%, suggesting that aggregating multi-level features at each scale can benefit spatial alignment and inference. In the second experiment, we utilize bilinear interpolation to replace RAM for upsampling low-resolution features, which decreases the accuracy by 1.5%. This result reveals that spatial misalignment is the key factor limiting scores fusion, and our RAM can effectively address this problem. Finally, we use concatenation operation followed by a 3×3 convolutional layer to fuse multi-scale features first, and then use a segmentation head to obtain the prediction. We conduct this experiment to compare the commonly used feature-level fusion with our in-network score-level fusion. Our ASFM reduces the computational complexity by 17.2 GFLOPs while improving the accuracy by 0.6%, which proves the effectiveness and efficiency of the adaptive score-level fusion mechanism.

4.2.6. Ablation Study for Booster Training Strategy

Table. 6 gives the performance comparison of MSL and MJS with and without OHEM, respectively. Compared to MSL, our MJS brings a 0.9% reduction in accuracy. This may be due to the fact that directly adding boundary supervision at each

Table 8: Experiments for the scale selection during the inference to further reduce model complexity. All ablations are trained on the complete architecture with four scales and discarded certain scales **only during the inference phase**. The results reaching the best trade-off are bolded.

Scale1	Scale2	Scale3	Scale4	mIoU (%)	Speed (FPS)	FLOPs (G)	Params (M)
✓	-	-	-	77.10	70.0	76.1	11.9
-	✓	-	-	77.94	64.4	81.3	12.2
-	-	✓	-	77.73	61.0	82.7	12.5
-	-	-	✓	77.01	57.9	82.8	12.7
-	✓	✓	✓	78.06	53.4	96.9	12.9
✓	-	✓	✓	77.93	53.6	96.9	12.9
✓	✓	-	✓	78.07	53.3	96.9	12.9
✓	✓	✓	-	78.10	55.1	96.7	12.8
✓	✓	✓	✓	78.16	52.2	101.6	13.0

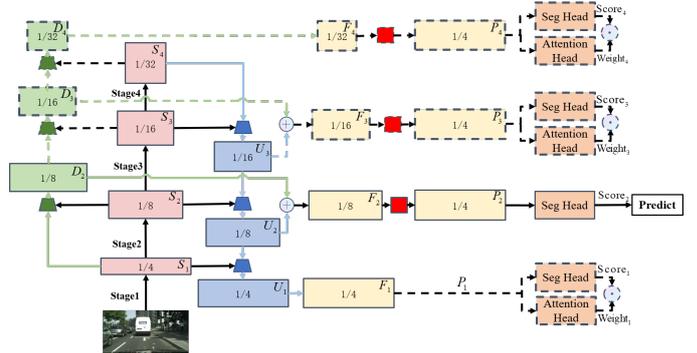


Figure 9: The illustration of one of the ablations in Table. 8. All dashed boxes are removable components that do not participate in the forward propagation to calculate the score map Scale2 during the inference process of MFARANet. This demonstrates a unique pruning method dedicated to the MFARANet, which is training on the whole network and removing certain scales during inference.

scale will lead to a more serious class imbalance problem, since boundary pixels only occupy a small part of the whole image. Boundary pixels are difficult training examples, so we employ OHEM to training on hard pixels selected by current loss to alleviate the class imbalance problem. After combing with OHEM, the accuracy of MJS + OHEM exceeds MSL + OHEM. The accuracy increase of MJS is 2.1%, while MSL only improves by 0.1%. This proves that our MJS is highly complementary to OHEM, and using them together maximizes the accuracy of our model. By adopting the booster training strategy, our model finally achieves a segmentation accuracy of 78.2% with real-time speed on the validation set of Cityscapes.

4.2.7. Ablation for Multi-scale Predictions

In Fig. 8, we compare the segmentation accuracy of different classes at each scale from the same trained model. From the results, Scale4 achieves higher segmentation accuracy than other scales on relative large objects marked in green. This due to the fact that the successive strided convolutional layers in DCNNs bring a large receptive field, which can reduce the intra-class confusion of large-scale objects. Scale2 and Scale3 have better accuracy on objects marked in blue and red because these middle or small-scale objects require a smaller receptive field and higher spatial resolution to segment details. Interestingly,

Table 9: Comparison with **real-time methods** on the Cityscapes dataset. In addition to using the whole picture as input to compute mIoU, we also provide single-scale inference accuracy on the crop size 1024×1024 . ‘MFARANet-Pruned’ indicates the pruning method specific to MFARANet, which is shown in Fig. 9 and discussed in Section. 4.2.8.

Name	mIoU(%)		Params(M)	Resolution
	Val	Test		
Speed-Orient:				
ENet [63]	-	58.3	0.4	360×640
ESPNet [64]	-	60.3	0.4	512×1024
BiSeNetV1-1 [54]	69.0	68.4	5.8	768×1536
ContextNet [65]	-	66.1	0.9	1024×2048
Fast-SCNN [66]	68.6	68.0	1.1	1024×2048
DFANetB [21]	-	67.1	4.8	1024×1024
EDANet [67]	-	67.3	0.7	512×1024
CGNet_M3N21 [68]	-	64.8	0.5	360×640
CIFReNet [69]	72.9	70.9	1.9	512×1024
NDNet [40]	-	65.7	1.1	1024×2048
MiniNet-V2 [70]	-	70.5	0.5	512×1024
AGLNet [71]	-	70.1	1.1	512×1024
SGCPNet2 [43]	-	69.5	0.6	768×1536
Good Trade-Off:				
SegNet [46]	-	57.0	29.5	360×640
ERFNet [7]	70.0	68.0	20.0	512×1024
ICNet [8]	-	69.5	26.5	1024×2048
BiSeNetV1-2 [54]	74.8	74.7	49.0	768×1536
SwiftNet-18 [12]	75.5	75.4	11.8	1024×2048
SwiftNet-18pyr [12]	74.4	75.1	12.9	1024×2048
ShelfNet [41]	-	74.8	23.5	512×512
BiSeNetV2-L [72]	75.8	75.3	-	512×1024
SFNet(DF1) [42]	-	74.5	9.0	1024×2048
STDC2-Seg75 [73]	77.0	76.8	12.5	768×1536
RGPNet [24]	74.1	-	17.8	1024×2048
FaPN2 [48]	75.6	75.0	12.6	768×1536
DMA-Net [22]	-	75.6	14.6	768×1536
MFARANet-Pruned	77.7 77.9	76.8 77.0	12.2	1024×1024 1024×2048
MFARANet	77.9 78.2	77.1 77.3	13.0	1024×1024 1024×2048

Scale1 has a smaller receptive field and higher spatial resolution than Scale2, but performs worse on small-scale objects. We attribute this to inputting very large-scale images of size 1024×2048 for inference, which may cause the receptive field of Scale1 to be too small for small-scale objects. It is worth noting that the final prediction achieves the highest segmentation accuracy in almost all classes. Combined with the visualization results in Fig. 7, the above results indicate that our MFARANet can achieve the goal of fusing multi-scale scores to favor objects of multiple scales.

Furthermore, we conduct several ablation experiments in Table. 7 on how multi-scale maps contribute to the final results. When removing the score map at a certain scale, we additionally add a segmentation head and a boundary head to calculate auxiliary loss, ensuring that all ablations can use the same loss MJS+OHem introduced in Section. 3.5. This also helps main-

Table 10: Comparison with **accuracy-oriented methods** on the Cityscapes test set. All models are trained using only the finely annotated dataset and the results are reported using **multi-scale testing**. FLOPs and FPS are computed on a single GTX 3090 GPU with an input size of 1024×1024 .

Name	FLOPs (G)↓	Speed (FPS)↑	Params (M)↓	mIoU (%)
FCN8s [30]	654.9	12.3	50.5	65.3
DeepLab-V2 [38]	1236.8	4.1	44.0	70.4
RefineNet [74]	979.1	9.8	118.4	73.6
SAC [75]	-	-	-	78.1
DepthSeg [76]	-	-	-	78.2
PSPNet [10]	696.2	8.5	65.7	78.4
ResNet38 [77]	2231.1	3.4	-	78.4
BiSeNet-V1 [54]	193.4	25.8	51.0	78.9
DFN [78]	1006.3	8.2	90.7	79.3
PSANet [79]	806.0	-	85.6	80.1
DANet [80]	1052.1	5.4	66.6	81.5
HRNet-V2 [15]	348.1	14.1	65.9	81.6
MFARANet-Pruned	81.3	64.4	12.2	78.5
MFARANet	101.6	52.2	13.0	78.7

tain the structural integrity of the MFAM and RAM. The results show that training on a single-scale map reduces the accuracy by 0.96%-1.65% with a promotion of only 4.7 FPS, while discarding one of the four score maps results in a decrease of 0.49%-1.04% with almost the same model complexity. Therefore, we can conclude that training on full-scale score maps is necessary as it achieves a better trade-off than the above ablations.

4.2.8. Ablation for Scale Selection

We conduct experiments in Table. 8 for the scale selection to further enhance the real-time performance. Since the results in Table. 7 have proved that training on full-scale score maps is necessary, we discard certain scales only during the inference phase from the trained complete architecture with four scales. Different with the experiments in Table. 7 that only discard the segmentation head and the attention head, the ablations in Table. 8 can remove all components that do not participate in the forward propagation of the network to calculate the score maps (illustrated in Fig. 9). From the results, inference on the single-scale score map Scale2 reaches the best trade-off, which is only 0.22% lower in the accuracy than the complete network with a promotion of 12.2 FPS. This can be seen as a pruning method dedicated to the MFARANet to reduce model complexity, which is training on the whole network and pruning during inference. The results of MFARANet-Pruned in Table. 9 can also prove the effectiveness of this pruning method on the Cityscapes test set.

4.3. Comparison With State-of-the-Arts on the Cityscapes

4.3.1. Accuracy Comparison

We compare our method with other real-time semantic segmentation methods in Table. 9. We roughly split them into two categories: the first one is speed-oriented, which has extremely low model complexity (FLOPs or Params); the other achieves

Table 11: Performance on CamVid. We use the whole picture as input to compute mIoU.

Name	Backbone	mIoU(%)	Params(M)
ENet [63]	-	51.3	0.4
SegNet [46]	-	55.6	29.5
BiSeNetV1-1 [54]	Xception-39	65.6	5.8
BiSeNetV1-2 [54]	ResNet-18	68.7	49.0
ICNet [8]	ResNet-50	67.1	26.5
DFANetA [21]	Xception-A	64.7	7.8
DFANetB [21]	Xception-B	59.3	4.8
SwiftNetRN-18 [12]	ResNet-18	72.6	11.8
CAS [81]	-	71.2	-
EDANet [67]	-	66.4	0.7
BiSeNet-V2 [72]	-	72.4	-
SFNet(DF2) [42]	DFNet-V2	70.4	10.5
CIFReNet [69]	Mobilenet-V2	64.5	1.9
NDNet45-FCN8-LF [40]	-	57.5	1.1
RGPNet [24]	-	66.9	17.8
CGNet_M3N21 [68]	-	65.6	0.5
MFARANet	ResNet-18	72.9	13.0

a good trade-off and can reach relatively high accuracy with moderate model complexity. Compared to these speed-oriented method, our network improves segmentation accuracy by at least 6.4% on the test dataset while maintaining real-time speed on a relatively high-resolution image of 1024×2048 . In particular, we outperform the well-known ENet by a large margin of 19.0%. As for the balance-oriented methods, they improve the accuracy as much as possible while keeping a real-time inference speed of over 24 FPS. SwiftNet and SFNet achieve close accuracy to our method, but both of them employ modified versions of Pyramid Pooling Module (PPM) [10] on top of their backbone to obtain multi-scale features, significantly improving their performance. Overall, without using additional modules (e.g., Atrous Spatial Pyramid Pooling [38], PPM) or their modified versions, our method with moderate model size achieves the highest segmentation accuracy among all real-time methods.

Table. 10 shows the comparison of accuracy and model complexity with accuracy-oriented methods, most of which use a very deep classification network (e.g., ResNet-101) or a well-designed large-scale network (e.g., HRNet) as the backbone. Our method shows a very competitive result with these state-of-the-art methods with much smaller model complexity. Compared to the most accurate method HRNet-V2, our network reduces the accuracy by less than 3.0% at 19.7% of the model size, 29.2% of the FLOPs and 3.7× FPS. Furthermore, our network reaches a higher segmentation accuracy with a much smaller model complexity than most methods, including the well-known DeepLab, RefineNet, PSPNet, etc. For example, our model outperforms PSPNet by 0.3% mIoU with 14.6% FLOPs, 19.8% Params and 6.2× FPS of it. The above results show that our network has a light-weight and effective decoder, and demonstrates this multi-scale parallel inference architecture is one way to achieve efficient semantic segmentation.

Table 12: Performance on PASCAL-Context. All results are reported using multi-scale testing. We also report the performance of our model with different backbones.

Name	Backbone	mIoU(%)	Params(M)
FCN-8s [30]	VGG	37.8	134.0
CRF-RNN [82]	-	39.3	-
ParseNet [83]	-	40.4	-
HO-CRF [84]	-	41.3	-
Piecewise [85]	-	43.3	-
VeryDeep [86]	-	44.5	-
DeepLab-v2 [38]	ResNet-101	45.7	43.9
Global Context [87]	ResNet-101	46.5	-
DenseDecoder [18]	ResNeXt-101	47.8	-
RefineNet [74]	ResNet-101	47.1	118.0
RefineNet [74]	ResNet-152	47.3	134.0
ShelfNet50 [41]	ResNet-50	45.6	38.7
ShelfNet101 [41]	ResNet-101	48.4	57.7
EncNet [14]	ResNet-50	49.2	35.5
EncNet [14]	ResNet-101	51.7	54.5
Our MFARANet	ResNet-18	46.7	13.0
	ResNet-50	49.6	25.7
	ResNet-101	50.9	44.7

4.3.2. Trade-off between Speed and Accuracy

Fig. 1 reflects the trade-off between FPS and accuracy, from which we have the following observations: 1) Our MFARANet and its pruned versions using different input size reach the highest accuracy among all real-time methods. 2) Only STDC2-Seg75 could reach competitive accuracy with our model (only 0.3% lower than us) on the test set at a similar inference speed, while lower than us by 0.9% on the valuation set (Table. 9). Additionally, our pruned-model has the same accuracy as STDC2-Seg75 on the test set, and reaches 0.7% higher accuracy on the valuation set with a speed increase of 12.1 FPS. 3) For methods having a faster inference speed, our MFARANet-pruned at 64.4 FPS achieves 1.2%-9.5% accuracy improvement on the test set. Thus, we can draw the conclusion that our framework exhibits a better trade-off than the compared methods.

4.4. Evaluation on Other Datasets

4.4.1. Performance on CamVid

CamVid is another dataset of road scene from a car perspective, which has much fewer training and testing samples than Cityscapes. We report the results on it in Table. 11 to prove the generality of our approach, from which we can draw similar conclusions as on the Cityscapes dataset. Compared to speed-oriented methods, our method shows a large increase in accuracy. For example, compared to ENet and CGNet, the accuracy of our method improves by 21.6% and 7.3%, respectively. Among methods aiming at good trade-off, we achieve the highest accuracy with a moderate model size.

4.4.2. Performance on PASCAL-Context

PASCAL-Context is a very challenging scene understanding dataset. Quantitative results are shown in Table. 12, and all

comparisons are picked from accuracy-oriented methods. Using ResNet-18 as the encoder, our method can outperform half of the accuracy-oriented methods in the table with very few parameters. By employing ResNet-50, the MFARANet can exceed the performance of all methods except EncNet using ResNet-101, including RefineNet adopting the very deep classification network ResNet-152 as the backbone. In particular, when using the same ResNet-50 as the encoder, our method improves the accuracy by 0.4% and 4% compared to EncNet and ShelfNet50, while having a much smaller model size (25.7M vs. 35.5M and 38.7M). The above results demonstrate the efficiency and effectiveness of our method.

5. Discussion

Our work conducts an in-depth and detailed study of the parallel inference architecture to make it suitable for real-time segmentation tasks. In this section, combined with the experimental results, we summarize the overall performance of MFARANet, and mainly discuss the motivation and substantiation of the three core components proposed for building our parallel inference network. As for the analysis of Multi-scale Joint Supervision and the proof that ASFM can favor objects of multiple scales, we provide them in Section. 4.2.6 and Section. 4.2.7 respectively.

The results in Table. 9 and Table. 11 show that our network achieves the highest accuracy with a moderate model complexity among all real-time methods on the Cityscapes and CamVid datasets. The trade-off comparison in Fig. 1 further prove that our model achieves the best balance between performance and model complexity. Besides, we achieve better or competitive performance with much smaller model complexity compared to accuracy-oriented methods in Table. 10 and Table. 12. The above results demonstrate the efficiency of the proposed multi-scale parallel inference network, and show its potential in combination with more advanced backbones for accuracy-oriented methods.

The MFAM, RAM, and ASFM are specially designed for parallel inference segmentation network. Firstly, Table. 2 proves that the MFAM is more suitable for segmentation than FPN because of aggregating multi-level features to each scale to provide hierarchical clues for inference. Secondly, by comparing RAM with bilinear interpolation and Straightforward Alignment, Table. 3 demonstrates that spatial misalignment is one of the key issues for pixel-level scores fusion and can be efficiently alleviated by our RAM. Finally, for ASFM, Table. 4 and Table. 7 show the effectiveness of adaptive fusion and the necessity of the prediction at each scale. Furthermore, the third row of Table. 5 proves that this adaptive score-level fusion is more effective and efficient than the feature-level fusion and single-scale prediction used by most segmentation networks [25, 19, 42].

Our network has some limitations that can be improved. The pixel-level attention mechanism in ASFM helps improve the performance, but it also increases the computational budget of the network. Although the pruning method shown in Fig. 9 and discussed in Section. 4.2.8 largely reduces model complexity, it is highly specific for our parallel inference architecture and

its generality need to be further studied. Therefore, designing a fast attention mechanism has the potential to further improve the speed and generality of our networks. Another future work is to build Transformer branches upon the MFAM to explore the combination of this parallel inference architecture with the self-attention mechanism [88].

6. Conclusion

In this paper, we present a Multi-level Feature Aggregation and Recursive Alignment Network (MFARANet) to perform multi-scale parallel inference in a single-pass network, a novel architecture for accurate real-time semantic segmentation. Specifically, our method consists of three core components: Multi-level Feature Aggregation Module (MFAM), Recursive Alignment Module (RAM) and Adaptive Scores Fusion Module (ASFM). MFAM is proposed to aggregate the hierarchical features of the encoder into four independent scales, which not only provides multi-level information for accurate segmentation, but also benefits the following spatial alignment operation. Combining the flow-based alignment module with the recursive upsampling architecture, our RAM can perform efficient and accurate spatial alignment between scaled score maps. Finally, we design ASFM to adaptively fuse multi-scale scores to generate the final prediction that favor objects of various scales. Experiments on three challenging datasets show that our MFARANet can provide a general and effective solution for real-time semantic segmentation.

Statements of Ethical Approval

We used three published datasets, so ethical approval was not required.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] J. Zhang, Y. Shi, J. Yang, Q. Guo, Kd-scfnets: Towards more accurate and lightweight salient object detection via knowledge distillation, *Neurocomputing* 572 (2024) 127206.
- [2] S. Chen, P. Sun, Y. Song, P. Luo, DiffusionDet: Diffusion model for object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19830–19843.
- [3] Y. Zhang, L. Feng, H. Shan, L. Yang, Z. Zhu, An aer-based spiking convolution neural network system for image classification with low latency and high energy efficiency, *Neurocomputing* 564 (2024) 126984.
- [4] K. I. Rashid, C. Yang, C. Huang, Fast-dsagcn: Enhancing semantic segmentation with multifaceted attention mechanisms, *Neurocomputing* (2024) 127625.
- [5] J. Chen, J. Lu, X. Zhu, L. Zhang, Generative semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7111–7120.

- [6] P. de Jorge, R. Volpi, P. H. Torr, G. Rogez, Reliability in semantic segmentation: Are we on the right track?, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7173–7182.
- [7] E. Romera, J. M. Alvarez, L. M. Bergasa, R. Arroyo, Erfnet: Efficient residual factorized convnet for real-time semantic segmentation, *IEEE Trans. Intell. Transp. Syst.* 19 (1) (2017) 263–272.
- [8] H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia, Icnnet for real-time semantic segmentation on high-resolution images, in: Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 405–420.
- [9] L.-C. Chen, Y. Yang, J. Wang, W. Xu, A. L. Yuille, Attention to scale: Scale-aware semantic image segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 3640–3649.
- [10] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2881–2890.
- [11] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 447–456.
- [12] M. Orsic, I. Kreso, P. Bevandic, S. Segvic, In defense of pre-trained image-net architectures for real-time semantic segmentation of road-driving images, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 12607–12616.
- [13] A. Tao, K. Sapra, B. Catanzaro, Hierarchical multi-scale attention for semantic segmentation, 2020, *arXiv:2005.10821*. [Online]. Available: <https://arxiv.org/abs/2005.10821>.
- [14] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, A. Agrawal, Context encoding for semantic segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7151–7160.
- [15] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* 43 (10) (2020) 3349–3364.
- [16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2117–2125.
- [17] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, H. Huang, Multi-scale context intertwining for semantic segmentation, in: Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 603–619.
- [18] P. Bilinski, V. Prisacariu, Dense decoder shortcut connections for single-pass semantic segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 6596–6605.
- [19] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, Unified perceptual parsing for scene understanding, in: Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 418–434.
- [20] N. Dvornik, K. Shmelkov, J. Mairal, C. Schmid, Blitznet: A real-time deep network for scene understanding, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 4154–4162.
- [21] H. Li, P. Xiong, H. Fan, J. Sun, Dfanet: Deep feature aggregation for real-time semantic segmentation, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 9522–9531.
- [22] X. Weng, Y. Yan, G. Dong, C. Shu, B. Wang, H. Wang, J. Zhang, Deep multi-branch aggregation network for real-time semantic segmentation in street scenes, *IEEE Trans. Intell. Transp. Syst.* (2022).
- [23] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 8759–8768.
- [24] E. Arani, S. Marzban, A. Pata, B. Zonooz, Rgpnet: A real-time general purpose semantic segmentation, in: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., 2021, pp. 3009–3018.
- [25] A. Kirillov, R. Girshick, K. He, P. Dollár, Panoptic feature pyramid networks, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 6399–6408.
- [26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 21–37.
- [27] B. Hu, J. Huang, Y. Liu, Y.-W. Tai, C.-K. Tang, Nerf-rpn: A general framework for object detection in nerfs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23528–23538.
- [28] R. Kaur, S. Singh, A comprehensive review of object detection with deep learning, *Digital Signal Processing* 132 (2023) 103812.
- [29] Z. Zou, K. Chen, Z. Shi, Y. Guo, J. Ye, Object detection in 20 years: A survey, *Proceedings of the IEEE* 111 (3) (2023) 257–276.
- [30] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3431–3440.
- [31] B. Shuai, Z. Zuo, B. Wang, G. Wang, Scene segmentation with dag-recurrent neural networks, *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (6) (2017) 1480–1493.
- [32] C. Peng, X. Zhang, G. Yu, G. Luo, J. Sun, Large kernel matters—improve semantic segmentation by global convolutional network, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4353–4361.
- [33] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, G. Wang, Semantic segmentation with context encoding and multi-path decoding, *IEEE Trans. Image Process.* 29 (2020) 3520–3533.
- [34] X. Ding, C. Shen, T. Zeng, Y. Peng, Sab net: A semantic attention boosting framework for semantic segmentation, *IEEE Trans. Neural Netw. Learn. Syst.* (2022).
- [35] D. Mazzini, Guided upsampling network for real-time semantic segmentation, 2018, *arXiv:1807.07466*. [Online]. Available: <http://arxiv.org/abs/1807.07466>.
- [36] D. Mazzini, R. Schettini, Spatial sampling network for fast scene understanding, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2019, pp. 0–0.
- [37] Z. Zhang, K. Zhang, Farsee-net: Real-time semantic segmentation by efficient multi-scale context aggregation and feature space super-resolution, in: Proc. IEEE Int. Conf. Rob Autom, IEEE, 2020, pp. 8411–8417.
- [38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (4) (2017) 834–848.
- [39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.-L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, *arXiv preprint arXiv:1412.7062* (2014).
- [40] Z. Yang, H. Yu, M. Feng, W. Sun, X. Lin, M. Sun, Z.-H. Mao, A. Mian, Small object augmentation of urban scenes for real-time semantic segmentation, *IEEE Trans. Image Process.* 29 (2020) 5175–5190.
- [41] J. Zhuang, J. Yang, L. Gu, N. Dvornik, Shelfnet for fast semantic segmentation, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), 2019, pp. 0–0.
- [42] X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, S. Tan, Y. Tong, Semantic flow for fast and accurate scene parsing, in: Proc. Eur. Conf. Comput. Vis. (ECCV), Springer, 2020, pp. 775–793.
- [43] S. Hao, Y. Zhou, Y. Guo, R. Hong, J. Cheng, M. Wang, Real-time semantic segmentation via spatial-detail guided context propagation, *IEEE Trans. Neural Netw. Learn. Syst.* (2022).
- [44] D. Lin, D. Shen, S. Shen, Y. Ji, D. Lischinski, D. Cohen-Or, H. Huang, Zigzagnet: Fusing top-down and bottom-up context for object segmentation, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7490–7499.
- [45] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 1520–1528.
- [46] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (12) (2017) 2481–2495.
- [47] Z. Huang, Y. Wei, X. Wang, W. Liu, T. S. Huang, H. Shi, Alignseg: Feature-aligned segmentation networks, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (1) (2021) 550–557.
- [48] S. Huang, Z. Lu, R. Cheng, C. He, Fapn: Feature-aligned pyramid network for dense image prediction, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 864–873.
- [49] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 764–773.
- [50] T. Takikawa, D. Acuna, V. Jampani, S. Fidler, Gated-scnn: Gated shape cnns for semantic segmentation, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 5229–5238.
- [51] S. Borse, Y. Wang, Y. Zhang, F. Porikli, Inverseform: A loss function for structured boundary-aware segmentation, in: Proc. IEEE/CVF Conf.

- Comput. Vis. Pattern Recognit., 2021, pp. 5901–5911.
- [52] J. Canny, A computational approach to edge detection, *IEEE Trans. Pattern Anal. Mach. Intell.* (6) (1986) 679–698.
- [53] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 761–769.
- [54] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Bisenet: Bilateral segmentation network for real-time semantic segmentation, in: *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.
- [55] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 3213–3223.
- [56] G. J. Brostow, J. Fauqueur, R. Cipolla, Semantic object classes in video: A high-definition ground truth database, *Pattern Recognit. Lett.* 30 (2) (2009) 88–97.
- [57] P. Sturgess, K. Alahari, L. Ladicky, P. H. Torr, Combining appearance and structure from motion features for road scene understanding, in: *Proc. BMVC*, 2009.
- [58] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, A. Yuille, The role of context for object detection and semantic segmentation in the wild, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 891–898.
- [59] J. Xie, B. Shuai, J.-F. Hu, J. Lin, W.-S. Zheng, Improving fast segmentation with teacher-student learning, 2018, *arXiv:1810.08476*. [Online]. Available: <http://arxiv.org/abs/1810.08476>.
- [60] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, G. Wang, Boundary-aware feature propagation for scene segmentation, in: *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6819–6829.
- [61] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, Y. Yan, Knowledge adaptation for efficient semantic segmentation, in: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 578–587.
- [62] M. Oršić, S. Šegvić, Efficient semantic segmentation with pyramidal fusion, *Pattern Recognit.* 110 (2021) 107611.
- [63] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, 2016, *arXiv:1606.02147*. [Online]. Available: <http://arxiv.org/abs/1606.02147>.
- [64] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, H. Hajishirzi, Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation, in: *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 552–568.
- [65] R. P. Poudel, U. Bonde, S. Liwicki, C. Zach, Contextnet: Exploring context and detail for semantic segmentation in real-time, 2018, *arXiv:1805.04554*. [Online]. Available: <http://arxiv.org/abs/1805.04554>.
- [66] R. P. Poudel, S. Liwicki, R. Cipolla, Fast-scnn: Fast semantic segmentation network, 2019, *arXiv:1902.04502*. [Online]. Available: <http://arxiv.org/abs/1902.04502>.
- [67] S.-Y. Lo, H.-M. Hang, S.-W. Chan, J.-J. Lin, Efficient dense modules of asymmetric convolution for real-time semantic segmentation, in: *Int. Conf. Multimed. Asia*, 2019, pp. 1–6.
- [68] T. Wu, S. Tang, R. Zhang, J. Cao, Y. Zhang, Cgnet: A light-weight context guided network for semantic segmentation, *IEEE Trans. Image Process.* 30 (2020) 1169–1179.
- [69] B. Jiang, W. Tu, C. Yang, J. Yuan, Context-integrated and feature-refined network for lightweight object parsing, *IEEE Trans. Image Process.* 29 (2020) 5079–5093.
- [70] I. Alonso, L. Riazuelo, A. C. Murillo, Mininet: An efficient semantic segmentation convnet for real-time robotic applications, *IEEE Trans. Robot.* 36 (4) (2020) 1340–1347.
- [71] Q. Zhou, Y. Wang, Y. Fan, X. Wu, S. Zhang, B. Kang, L. J. Latecki, Aglnet: Towards real-time semantic segmentation of self-driving images via attention-guided lightweight network, *Appl. Soft. Comput.* 96 (2020) 106682.
- [72] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, N. Sang, Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation, *Int. J. Comput. Vis.* 129 (11) (2021) 3051–3068.
- [73] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, X. Wei, Rethinking bisenet for real-time semantic segmentation, in: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9716–9725.
- [74] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1925–1934.
- [75] R. Zhang, S. Tang, Y. Zhang, J. Li, S. Yan, Scale-adaptive convolutions for scene parsing, in: *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2031–2039.
- [76] S. Kong, C. C. Fowlkes, Recurrent scene parsing with perspective understanding in the loop, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 956–965.
- [77] Z. Wu, C. Shen, A. Van Den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition, *Pattern Recognit.* 90 (2019) 119–133.
- [78] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Learning a discriminative feature network for semantic segmentation, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1857–1866.
- [79] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, J. Jia, Psanet: Point-wise spatial attention network for scene parsing, in: *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 267–283.
- [80] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
- [81] Y. Zhang, Z. Qiu, J. Liu, T. Yao, D. Liu, T. Mei, Customizable architecture search for semantic segmentation, in: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11641–11650.
- [82] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. H. Torr, Conditional random fields as recurrent neural networks, in: *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1529–1537.
- [83] W. Liu, A. Rabinovich, A. C. Berg, Parsenet: Looking wider to see better, 2015, *arXiv:1506.04579*. [Online]. Available: <http://arxiv.org/abs/1506.04579>.
- [84] A. Arnab, S. Jayasumana, S. Zheng, P. H. Torr, Higher order conditional random fields in deep neural networks, in: *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Springer, 2016, pp. 524–540.
- [85] G. Lin, C. Shen, A. Van Den Hengel, I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, in: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 3194–3203.
- [86] Z. Wu, C. Shen, A. v. d. Hengel, Bridging category-level and instance-level semantic image segmentation, 2016, *arXiv:1605.06885*. [Online]. Available: <http://arxiv.org/abs/1605.06885>.
- [87] W.-C. Hung, Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, M.-H. Yang, Scene parsing with global context embedding, in: *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2631–2639.
- [88] X. Li, H. Ding, W. Zhang, H. Yuan, J. Pang, G. Cheng, K. Chen, Z. Liu, C. C. Loy, Transformer-based visual segmentation: A survey, *arXiv preprint arXiv:2304.09854* (2023).